A Trial Bank Model for the Publication of Clinical Trials

Ida Sim, M.D., and Glenn Rennels, M.D. Ph.D.

Section on Medical Informatics, Stanford University School of Medicine
Stanford University, Stanford CA
sim@camis.stanford.edu

Abstract. *Clinical trials constitute one of the main sources of medical knowledge, yet trial reports are difficult to find, read, and apply to clinical care. Reasons for these difficulties include the lack of a common, standardized, structure for trial reports; the restricted length of reports; and limited computer support for use of the literature. We propose a new model of reporting clinical trials, in which trials are published as both prose commentary and as data in electronic "trial banks." The prose will allow authors to discuss their trials in writing; the electronic database will allow readers easy access to well-defined data about the trials. We are developing a formal conceptual model of the clinical trials domain for integrating the use of multiple trial banks. We will then focus on validating this conceptual model with clinical literature users.*

INTRODUCTION

Our society invests heavily in clinical research, but researchers and practitioners cannot easily find, synthesize, or interpret the results of this research. Traditional academic publishing is already undergoing major changes in anticipation of the digital age. Now is the appropriate time to redesign how we publish clinical trial reports so that new information can be incorporated quickly and effectively into clinical decision making.

We first summarize the current problems in using the clinical trials literature, discuss some proposed solutions, and report on a needs analysis we conducted on a target group of literature users. We then present our proposal that clinical trials be reported into electronic databases, or "trial banks," as part of the publication of trial results. We conclude with some practical considerations for implementation of this proposal.

CURRENT PROBLEMS

Much has been published on deficiencies in the clinical literature's indexing (1), retrieval (2), quality (3), reporting (4), dissemination (5), and interpretation and clinical application (6). Medline searches identify as few as 48% of the relevant articles that could be found by hand-searching (1); one study estimated that 50% of articles use incorrect statistical methods (3); despite strong evidence in the literature that aspirin can reduce myocardial infarction mortality, many practitioners are unaware of this new knowledge or are not changing their practice to conform with it (5). Others have presented approaches to these difficulties, but while each approach successfully addresses some of the problems, none offer a comprehensive solution.

1) **Structured and standardized reporting.** The standardization of abstracts (7) and, recently, of trial reports (4), has helped to improve the clarity, precision, and completeness of trial reports (8). This standardization has made interpretation and use of the reports easier, but indexing, retrieval, and dissemination problems are not addressed.

2) **Electronic publication.** Electronically published trial reports may be longer and more complete. They can also be better disseminated, but problems of retrieval, quality, interpretation, and use would still exist.

3) **Trial registries.** Trial registries list clinical trials that are planned, ongoing, or completed. Registries can promote patient enrollment into trials, and can help researchers coordinate the planning of trials. Current registries are not easily or widely accessible (9).

4) **Systematic reviews of the literature.** The Cochrane Collaboration (10) uses rigorous methodology to prepare and maintain "systematic reviews of the effects of health care" and will disseminate these reviews electronically. While Cochrane reviews are valuable for their sound interpretations of the literature, this approach does not address basic problems in the indexing, retrieval, quality, and reporting of individual trials.

5) **Expert systems.** Systems such as Roundsman (11) and THOMAS (12) have attempted to assist clinicians with direct use of clinical trial results. Practical use of such systems is severely limited by the need to manually update their knowledge bases with new trials.

Each of the above approaches has its merits, but the number and variety of the problems suggests that a comprehensive solution is needed. Because solutions

0195–4210/95/$5.00 © 1995 AMIA, Inc.

are often better if they are grounded in the needs of those who have the problem, we analyzed the needs of a target group of clinical literature users.

Needs Analysis

Users of the clinical literature include health care providers, students, and policy makers; researchers; biostatisticians and meta-analysts; journalists; and the general public. Of these user groups, meta-analysts make the most intensive and rigorous demands on the literature. Meta-analysts pool data from trials of similar design to achieve higher statistical power. For the data pooling to be valid, they must retrieve all relevant trials and the data must be accurate, complete and unambiguous. These needs are central to use of the trials literature, regardless of any controversy about the validity of meta-analytic methodology.

Since meta-analysts face all the problems that other users may encounter, we chose meta-analysts to be our target user group for grounding any proposal to promote more effective use of the clinical trials literature. We asked six practicing meta-analysts to describe the information management problems they have experienced in meta-analysis. There were four major themes of dissatisfaction:

1) Medline retrieval misses relevant studies because many concepts are unreliably keyworded (e.g., study design), or are not keyworded at all (e.g., sample size). Trials that are unpublished or are ongoing are not in Medline. Meta-analysis of only published trials may be biased, since published trials are more likely to have non-negative results.

2) Trial designs are inadequately reported, making it difficult to judge trial validity.

3) Trial results are often ambiguously and incompletely reported, or the data are internally inconsistent. In trials addressing similar clinical questions, different outcomes are measured and may be reported in differing formats which preclude data pooling.

4) Extracting information from trial reports is time and labor intensive, and is prone to error. Conceptually related data are not necessarily reported together. Data have to be manually entered into computer programs for further analysis.

Many of the above problems stem from the absence of a well-defined, standardized, structure for trial reports. We are also unable to use the computer to directly manage the information in trial reports, whether for information retrieval or for analysis. The reporting of trials in natural language impedes any comprehensive solution to these problems. Natural language is not a good medium for well-defined, structured, reporting, whereas electronic databases are. Furthermore, natural language cannot be understood by computers, whereas databases can.

THE TRIAL BANK MODEL

We propose that trials be published concurrently in two forms: in electronic database form, and in traditional prose form. This mode of publishing is called electronic data publishing (13). The electronic database will provide clear and declarative semantics, computer accessibility, and reusability of the data for multiple purposes. The prose commentary will provide the readability and expressiveness of natural language.

Electronic data publishing already exists. Several major molecular biology journals require that authors submit their genomic sequence data directly into GenBank, a database of nucleotide sequences, before their manuscripts can be accepted for publication. When the article is published in traditional paper form, its accession number to the GenBank database is appended so that readers can have direct computer access to the article's sequence data.

In clinical trials publishing, medical journals could require that authors submit their trials into a trial bank in addition to submitting prose manuscripts. When a manuscript is published, the accession number to the corresponding trial bank entry will allow readers to have direct computer access to detailed and structured information about the trial's design, execution, and results. The reader can then use this data for further analysis, or expert systems can download this data for use in decision support.

For example, if an electronic database contained information on at least a trial's population, sample size, intervention and primary outcome, a query such as the following can be executed : "get all trials with a sample size over 100 that look at mortality in post-heart attack patients who are given aspirin." The prose discussions and the complete, structured, data of the resulting trials would now be directly available for further querying or analysis. No existing general system has the framework to offer this broad functionality.

System Architecture

The trial bank system architecture has three major components: 1) the trial banks, of which there will likely be many worldwide, 2) the data structures of each trial bank, and 3) the shared clinical trials ontology (Figure 1).

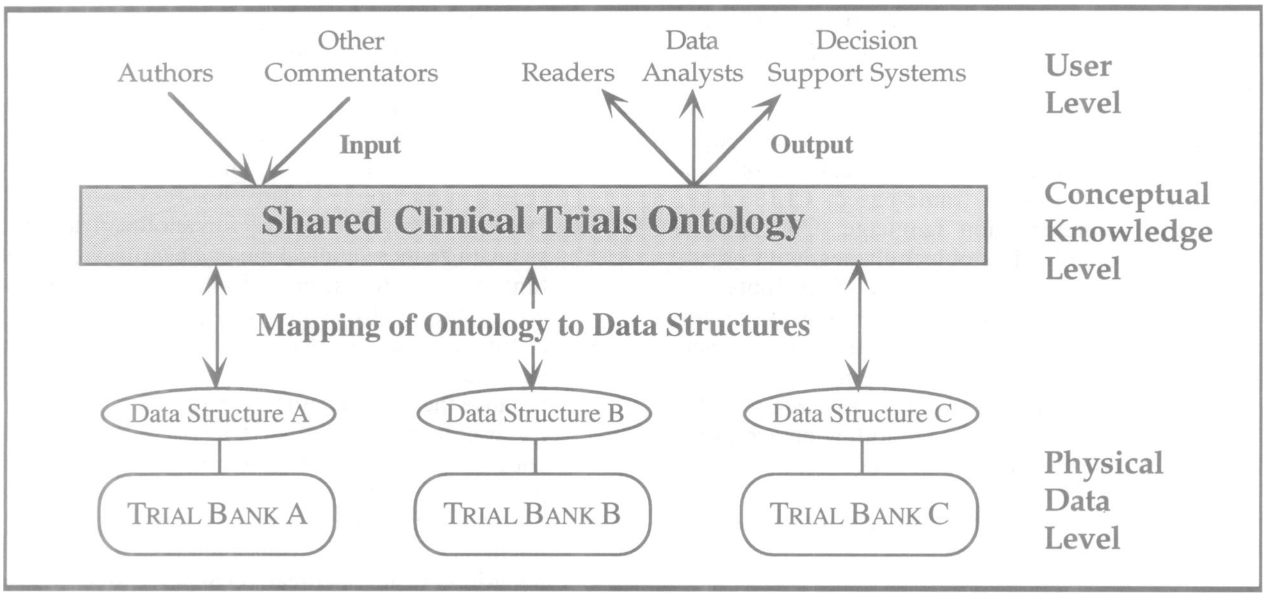

Figure 1. Trial banks may contain information about trials in all stages of execution, from planned to completely reported. There will likely be many different trial banks, each with their own data structure. Authors of trial reports will write into trial banks using authoring software, as will authors of letters to the editors and other commentary. Queries can be submitted automatically to multiple trial banks by expressing the query in the shared ontology and then mapping the query to each trial bank's data structure. Users can write into the trial banks or query them by using any application program that can use the shared ontology. Because this ontology will be shared, it should be agreed on by those who will be using it. Our ontology already incorporates concepts suggested by experts (4) on trials reporting, and we will work to build a consensus as the ontology evolves.

In the artificial intelligence field, an ontology is defined as an "explicit specification of . . .the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them. [An ontology] is an abstract simplified view of the world that we wish to represent for some purpose." (14) The clinical trials ontology is thus an abstract, conceptual view of clinical trials for the purpose of communicating, to both people and machines, any trial's design, execution, and results. The shared ontology will allow users to interact with many different trial banks at the conceptual rather than the data level. For example, a query expressed in the shared ontology could be mapped automatically to the data structure of each individual trial bank, and then executed without the user having to know anything about any data structures. In the latter scenario, the shared ontology will act as a global database schema for the interoperation of multiple databases.

Developing the Ontology

There is no formal methodology for developing ontologies. The first step is to learn about the domain; the next step is to conceptualize the structure of the concepts in the domain for a particular task. For the clinical trials domain, the central task is to synthesize data from many trials to arrive at the best evidence-based knowledge for clinical use. Meta-analysis is one method for performing this task, so we chose to conceptualize the domain for the purpose of meta-analysis. As stated before, meta-analysts have all the needs, and more, of other trial report users so the conceptualization should also be applicable for other less demanding purposes.

Conceptualization. One of us joined a large meta-analysis project in order to conceptualize as meta-analysts do. We iteratively refined our understanding of the clinical trial concepts we needed, and how these concepts related to each other. For example, the percent of patients followed-up for assessment of outcomes is often conceptualized as a single attribute of an entire trial. However, we realized that most trials assess more than one outcome, and the follow-up can be different for each outcome. Therefore, trials have a follow-up for *each* outcome assessed. These conceptualizations were reflected in the design of an abstraction form for capturing all the relevant details from trial reports. Five meta-analysts participated in the conceptualization.

Modeling. For a conceptualization to become an ontology in the sense of an "explicit specification" of concepts, it should be expressed in a formal knowledge representation language. A formal language forces explicitness and clarity. This explicitness, or declarativeness, is useful for clear communication and for highlighting when, for example, two databases have different meanings for the same concept.

We used the Protégé-II ontology editor (15) to begin to formalize our conceptualization because it offers a graphical interface to a frame-based, CLIPS-like, knowledge representation language. Concepts in Protégé-II are defined as object classes, with object attributes as slots, and attributes of attributes as facets. The "outcome" concept discussed above, for example, would be represented as in Figure 2: each outcome is conceptually composed of the definition of the endpoint, how that endpoint was assessed, who it was assessed on and when, the follow-up achieved, and finally, the value of the outcome.

Our ontology now has 158 classes, which can be classified into two major types.

1) **Generic classes.** These classes represent fundamental concepts about the world, such as time and causality. Many of our generic classes were borrowed from the T-Helper ontology (16), which is a clinical trials ontology for monitoring patients in AIDS trials. Reusing the T-Helper ontology saved us much time and effort in modeling these complex concepts.

2) **Clinical trial classes.** These classes represent concepts about clinical trials and form the majority of the classes in our ontology. Examples include the classes 'outcome,' 'randomization,' and 'population.' New concepts such as allocation concealment (whether "the intervention assignment schedule [was concealed] from participants and clinicians until recruitment was complete and irrevocable" (4)) are explicitly modeled as relationships among populations, interventions, trial executors and time. Such declarative semantics, once agreed on by the user community, should enable clearer communication of concepts within the field.

```
(defclass outcome
    (slot endpoint-definition (type symbol)
      (allowed-classes endpoint))
    (slot endpoint-assessment (type symbol)
      (allowed-classes assessment-method))
    (slot assessed-population (type symbol)
      (allowed-classes treatment-population)
    (slot follow-up (type float)
      (cardinality single))
    (slot outcome-value (type symbol)
      (allowed-classes discrete-value
       continuous-value categorical-value
       count-value))
)
```

Figure 2. Example of a Protégé-II class definition.

The clinical trial classes will hold actual values for particular trials. For example, the clinical trial class 'inclusion-criteria' may have the value "Ejection fraction less than 40%" for one trial, and "Pregnant, first trimester" for another trial. To standardize these clinical medicine terms would be tantamount to standardizing the medical vocabulary. However, we postulate that a common, standardized medical vocabulary is not a necessary component of the trial bank model because the core benefits of the model come from representing the clinical trial as a conceptual entity, rather than from representing the medicine that underlies the trials. This key distinction between representing trials versus representing medicine allows the trials ontology to be valid regardless of the medical vocabulary used. Thus, the class 'inclusion-criteria' is valid within the trials ontology irrespective of how it is filled in.

We are currently developing our ontology to accommodate trials of all experimental designs and to represent all the concepts needed for the task of meta-analysis. We will then evaluate the ontology by first using it to describe an assortment of trials, and then meta-analyzing the trials. We plan to use the MeSH clinical medicine vocabulary as an adjunct to the trials ontology for development purposes.

A drawback of modeling ontologies in Protégé-II is that Protégé-II cannot express disjoint classes, functions or axioms. We plan to use Ontolingua (14) as our eventual knowledge representation language because Ontolingua can express all of first-order logic and because it is designed for knowledge sharing. We plan to be compatible with any emerging ontology and database interoperation standards.

Practical and technical considerations

Several practical and technical considerations must be addressed if trial banks are to exist. Publishers will want to retain their economic and proprietary role, and authors must retain their share of control. These concerns can be assuaged if publishers maintain their own trial banks, so that they can charge for access and can have their editors remain as the arbiters of quality for their publications. Other, possibly competing, trial banks may also be maintained by independent organizations and governments. Authors can retain their control because data that are currently unavailable, such as individual patient data, can remain unavailable. In return for being compelled to submit to a trial bank, authors will be helped by authoring software to accurately and completely report their trials. Because this system starts the data acquisition process with the authors, no additional labor is needed to translate trial reports from one format to another. Trial banks will thus be maintained as a byproduct of the business of academic medical publishing. Other difficult and important issues concerning intellectual property rights, payment for electronic documents, and control of research data remain unresolved.

Technical considerations include the still unproven use of ontologies for sharing information at the conceptual knowledge level. Automatic translation of ontologies to database structures would do much to facilitate the practical use of ontologies. Issues in ontological standardization, sharing, and integration also remain open.

CONCLUSIONS

Difficulties in using the clinical literature have been well-documented. Current approaches to these difficulties present only partial solutions, while a comprehensive solution is impeded by the reporting of trials in natural language. Therefore, we propose that trials be reported into databases that have well-defined and common data structures, instead of being reported only in natural language. We contend that the trials will be easier to identify and retrieve, will be more completely and accurately reported, and will be more easily accessible to computers for analysis and decision support. Our current work focuses on the development of a conceptual model of the clinical trials domain, and on validating this model with the clinical trials research community.

Acknowledgments. This work was partially supported by grant HS 08362 from the Agency for Health Care Policy and Research, Rockville, MD.

References

1. Dickersin K, Higgins K, Meinert CL. Identification of meta-analyses. The need for standard terminology. Control Clin Trials 1990;11(1):52-66.

2. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in Medline. Journal of the American Medical Informatics Association 1994;1(6):447-458.

3. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation 1980;61(1):1-7.

4. The Structured Reporting of Trials Group. A proposal for structured reporting of randomized controlled trials. JAMA 1994;272(24):1926-31.

5. Hennekens C, Jonas MA, Buring JE. The benefits of aspirin in acute myocardial infarction. Still a well-kept secret in the United States. Arch Intern Med 1994;154(1):37-9.

6. Evidence-based medicine Working Group. A new approach to teaching the practice of medicine. JAMA 1992;268(17):2420-5.

7. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. Ann Intern Med 1990;113(1):69-76.

8. Rennie D. Reporting randomized controlled trials. An experiment and a call for responses from readers. JAMA 1994;273(13):1054-5.

9. Easterbrook PJ. Directory of registries of clinical trials. Stat Med 1992;11(3):345-59.

10. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. Bmj 1994;309(6958):862-5.

11. Rennels G. A computational model of reasoning from the clinical literature. Berlin: Springer Verlag, 1987.

12. Lehmann, H. A bayesian computer-based approach to the physician's use of the clinical research literature. PhD Thesis in Medical Informatics, Stanford University, 1991.

13. Cinkosky MJ, Fickett JW, Gilna P, Burks C. Electronic data publishing and GenBank. Science 1991;252(5010):1273-7.

14. Gruber T. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 1993;5(2):199-220.

15. Gennari J. A brief guide to MAITRE and MODEL: An ontology editor and a frame-based knowledge representation language. Technical Report KSL-93-45, Knowledge Systems Lab, Stanford, CA, 1993.

16. Tu SW, Shortliffe EH, Gennari J, Shahar Y, Musen MA. Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: Application of Protege-II to protocol-based decision support. Technical Report KSL-94-22, Knowledge Systems Lab, Stanford, CA, 1994.